Computational problem

Huy Nguyen

February 2016

1 Introduction

By using the event-driven approach for studying gene block evolution in bacteria, it allows us to determines the combinations of possible events to transition between the tree nodes. In this paper, I will provide a parsimony approach to reconstruct the ancestral gene blocks given the phylogenetic tree and leaf nodes

2 Assumption

1. Phylogenetic Tree:
   - Our phylogenetic tree is given, it is binary and rooted.
   - Leaves are populated by orthoblocks. At least one leaf has a reference operon.
   - The model is agnostic to gene order.
2. Relationship between parent nodes and their children:
   - Given a parent gene blocks, its children gene blocks can't have any gene that is not in the parent gene blocks. This is a quite unrealistic. I need this to buil the initial set. However, by providing some correction mechanism (reduction), I will remove this assumption
   - There are 3 types of events that can happen from a parent to a child [1]:
     - Split : If two genes in one taxon are neighboring and their homologs in the other taxon are not, then that is defined as a single split event. The distance is the minimal number of split events identified between the compared genomes.
     - Deletion : A gene exists in the operon in the one taxon, but its homolog cannot be found in an orthoblock in another taxon. Note that the definition of homolog, e-value 10^{-10} is strict, and may result in false negatives. The deletion distance is the number of deletion events identified between the compared target genomes.
     - Duplication : A duplication event is defined as having gene j in a gene block in the source genome, and homologous genes (j',j") in the homologous block in the target genome. The duplication distance is the number of duplication events counted between the source and target genomes. The duplication has to occur in a gene block to be tallied.
   - Multiple events from parent to children are possible.
   - Events are treated as independent. (need more precise study)

3 Event-Based Distance

The distance between an internal node and its child is defined as below [1].

- Split distance d_s is the absolute difference in the number of relevant gene blocks between the two taxa. Example: for the reference gene block with genes (abcdefg) Genome A has blocks ((abc),(defg)) and genome B has ((abc),(de),(fg)). Therefore, d_s(A,B) = $\left|2-3\right| = 1$

- Duplication distance d_u is the pairwise count of duplications between two orthoblocks. Example: we have a reference gene block (abcde). Now, for genomes A and B the orthoblocks are A=((abd)) B=((abbcc)). Gene A_b is duplicated in genome B, thus a duplication distance d_u(A,B) of 1. Gene c generates a distance of one deletion (see below) and one duplication. This is because the most parsimonious explanation is

[1] An event-driven approach for studying gene block evolution in bacteria, Iddo Friedberg

that the most recent ancestor for A and B may have had one copy of c, thus generating a duplication in one lineage, and a deletion in another. Because gene d exists only in the reference genome, it has no bearing on the event-based distance between the homologous gene blocks A and B.

- Deletion distance d_d is the difference in number of orthologs that are in the homologous gene blocks of the genome of one organism, or the other, but not in both.

4 Problem

Given our model assumption and the event-based distance, the computation question is to reconstruct ancestral gene blocks with the fewest number of events distance to its closest children.

5 Approach

I will append the transitional rationale in here